日本語作文・原稿チェックシステムの構築
—表記チェッカーの開発—

長谷川守寿

要　旨

本稿は、日本語を学習する学生の作文・原稿に対して、表記がおかしいと思われる部分を表示するチェックシステムの構築法と、そのシステムの評価についての考察である。学習者は、日常で触れたことのある語彙や学習した文法・語彙を使って作文を行う。これから提案するチェックシステムは、初級・中級の学習者が書く作文・原稿の中で、特に表記の誤りに対応できることを目指したものである。大きな特徴は学習者の学習状況に合わせて、文法・語彙を制限することによって、効率的な処理を行うことを目指した点である。本稿では、このシステムを用いることで、学習者が犯す様々な誤りのうち、表記や活用などのチェックを行うことが可能であることを、システム作成の手順と実行例をもとに明らかにする。

【キーワード】　表記チェックシステム、茶筌、語彙、連接規則

For a Development of Diagnostic System of Japanese Composition : representation checker

Hasegawa, Morihisa

Abstract

This paper reports on a development of a representation checker and its evaluation. A support system for Japanese written compositions has been developed for beginners, using Chasen. This is a Japanese morphology analysis system for Windows PCs. When a learner selects a lesson number he/she learned, this system reconstructs the dictionary and the connective rules that are thought to have been learned. When a learner inputs compositions, the system diagnoses the sentences, based on the learned dictionary and the connective rules. This shows the user where the words in compositions are not well-formed or thought to be mistakes.

I propose to construct the Japanese checker based on a student's profile of learning. The system is developed as an application of natural language processing.

1．目的

日本語の作文の授業では、留学生に作文を書いてもらい、教師が添削することを繰り返すのが通例である。そのため教師の添削を支援したり、自動的に添削を行うシステムのためには、コンピュータを用いた自動的な日本語チェッカーが有効であると考える。

従来の日本語作文チェックシステムの問題点として、これらのシステムが対象としている作文は中・上級者の書いたものが中心であり、初級学習者の作文を対象としたシステムは管見では見あたらない。そこで本研究では、形態素解析システムと、制限が加えられた辞書・文法規則を持つ初級学習者の作文に対応できるシステムを構築した。

本稿の目的は、初級日本語学習者の作文チェックを目指したシステムの構築法と、それに沿って作られたシステムによる作文チェックの評価について考察を加えることである。

2．先行研究

掛川他（2000）は、日本語学習者の作文に対して、正誤だけでなく、学習者がどのように誤っているのか、またどう直すべきかをコメントしたり、指導を行うことを目的とした診断システムである。学習者は具体的な場面設定と意味的係り関係に関する情報、前後の文脈、使用できる語のリストを制約として与えられた上で作文を行うこととなる。診断は学習者に与えられた制約に対応した別の制約を利用して行われ、語の過不足、活用、接続辞の誤り、余分な係りの可能性、係りの不足や障害、交差係り、また状況依存の表現における不適切さなどの診断が可能となっている。しかし、問題点として、自由作文に対応していないことが挙げられる。このシステムは特定の語の用法の学習には適していると思われるが、作文の授業で使用するには、制約を事前にたくさん用意する必要がある。

山本他（2000）は、入力される文の自由度を上げることを目指し、N-gramモデルに基づく統計的言語モデルを用いた日本語チェッカーを提示している。このシステムでは辞書はあらかじめ用意せず、文字トライグラムモデルで確率を計算して行っており、このシステムで行った「て形」の促音脱落・挿入誤りについては、十分実用的な結果が得られている。しかし、言語モデルとなっているのは新聞記事であり、留学生の作文はこれとはかなり異なる表記や内容を持つと考えられる。そのため、このモデルを直接作文チェッカーとして適用するには多くの課題があると思われる。よってどのような文章を言語モデルとすべきか問題が残る。

3．システム

3．1　システムの特色

本システムの大きな特徴は、あらかじめ全ての可能な文に対応できることを想定した文法項目・辞書を用意するのではなく、学習者に合わせた文法項目・辞書を随時用意する点にある。統計的手法を用いたシステムが主流である中、このような方法を用いたシステムは管見では見あ

たらない。本稿で提案するシステムの中心部となるのが、茶筌[1]という日本語形態素解析システムである。茶筌の特徴は、利用者が語彙や文法情報を自由に定義できる点である。文法情報とは、使用する品詞分類、活用形名、どの語とどの語がつながるかを記述した連接規則などからなる。本研究では、課ごとで学ぶ語彙、文法情報を特定し、茶筌で使用する語彙、文法情報（特に連接規則）を制限する。使用者は、どの課まで学習したかをあらかじめ登録することで、入力した作文に対して、学習した課までの語彙、文法情報のチェックが加えられることになる。

本システムの利用目的は、学習者の学習履歴に従い、文法情報や語彙を制限したチェッカーを使用することで、正しく形態素解析できない部分、すなわち誤りである可能性が高いため、何らかの修正が必要な部分を表示し、学習者に教科書等を基に再考をうながし、学習者自身の力で修正させることにある。

このシステムの構想には、学習者は作文を行う際に、未習の文法項目は正しく使用することは出来ないであろうということ、さらに学習者は誤りの箇所を指摘されれば、自分の力でどうにか修正できるのではないかという予測がある。このような考え方は、システム評価において、より多くの誤りの可能性に触れさせることが望ましいという本稿の立場にも関わる。そういった学生の持つ能力に対する見方が正しければ、このようなシステムは十分機能するのではないかと思われる。

本システムは、これだけで独立するのではなく、図１で示されるような格文法等を使用した作文診断システムの前処理を行うものとして位置付けられる。なぜなら、初級者の作文をそのまま作文診断システムの入力としても、誤表記がたくさん含まれるため正しい診断が行われるとは考えられないためである。そしてこれが終了した段階で、作文診断システムへ進むことを想定する。また、単独で使用する場合には、本システムは完全に教師の役割を代用するものではなく、教師の手間を減らすフィルターのような存在と位置付け、教師からのフォローが必須と考える。

図１．想定するシステム構成

３．２　処理の流れ

本研究では、図２に示すような処理の流れを想定する。作文を入力すると、語彙・連接規則が制限された茶筌によって形態素解析され、その結果、「未定義語」とされたものを含む文を表示する。これを見て学習者はさらに作文を修正し入力する、という手順を繰り返すことを考えている。

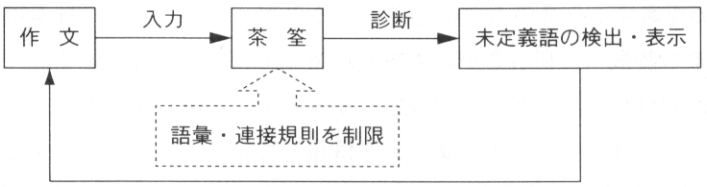

図２．処理の流れ

茶筌で形態素解析を行うと、語として認識されなかった部分は未定義語として解析される。本研究では、この未定義語を、制限された語彙・連接規則を持つシステムから語と認識されなかった文字（列）、すなわち誤りである可能性が高い部分として、一文単位で表示する。実際には、図３のような画面が表示されることを想定し、下線部が未定義語となる。

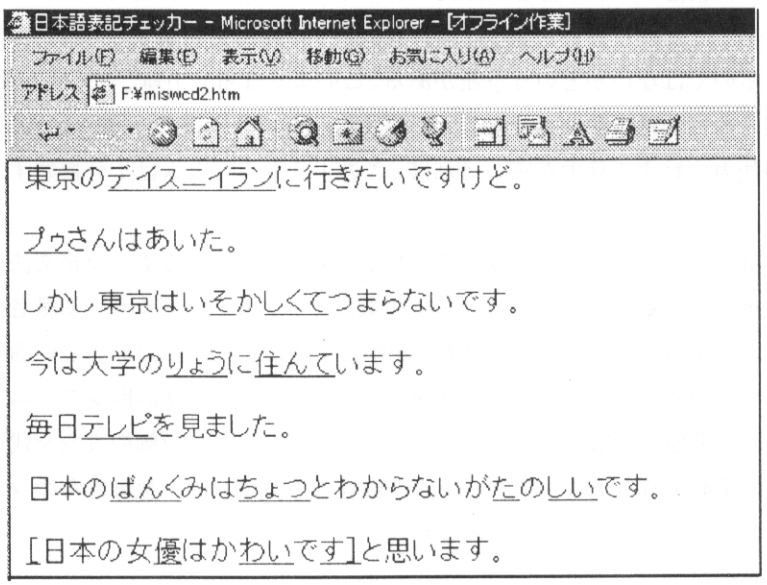

図３．実際の画面例

４．システム作成の手順

上記の考えに基づいたシステムを作成する。作成するシステムは全ての課に対応したシステムではなく、上記の考えを評価するために、ある課までの作文に限定した部分的なシステムである。以下は、システム作成のための手順である。

このシステムの利用者は、SFJ（Situational Functional Japanese）[2] を使用している学習者とする。今回、本システムの評価に使用する作文は、筑波大学留学生センター開講の日本語150という補講クラスで学期末試験の一つとして書かれたものであり、１・２学期分で計24人分、405文である。学生はその時点で12課までを授業で学習している。

この学習者の作文に対応させるために、12課までの語彙・連接規則に制限したシステムを作成する。以後、2種類の茶筌について言及するので、通常の茶筌をそのまま『茶筌』、12課までの内容に制限を加えた茶筌を特に『12茶』と呼ぶこととする。語彙は、SFJのDRILLSとNOTESの各巻から全て抜き出し[3]、品詞別に辞書に登録する。学習者が作文を書くときには、通常漢字で表記される語彙もいろいろな表記で書かれることが予想されるので、辞書登録に際し、ひらがな表記・漢字表記・混合表記を用意した。例えば「来年」という語の場合、(1) のような形式で、辞書に登録した。

(1) (名詞 (時相名詞 ((見出し語　来年　らいねん　らい年) (読み　らいねん))))

連接規則とは、ある表現にどのような表現が続くかを形式化したものである。このシステムでは、NOTES内のGrammar Notesに含まれる文の連接規則を特定し、茶筌で使用する連接規則を12課までの文法項目に制限した。特定の際には、茶筌で形態素解析を行い、文を形成するのに必要となる連接規則を決定した。(2) が実際に制限を加えた連接規則の例である。茶筌では、セミコロン“;”はないが、(2) のようにセミコロンを加えることにより、12茶ではセミコロンの後の括弧内の形態素構造[4]は接続されないようになる。

(2) (((動詞　＊　＊　基本連用形) ; (接尾辞　動詞性接尾辞　＊　基本連用形))
((接尾辞　形容詞性述語接尾辞　イ形容詞アウオ段　＊　たい)
; (接尾辞　形容詞性述語接尾辞　イ形容詞アウオ段　＊　やすい)
; (接尾辞　形容詞性述語接尾辞　イ形容詞アウオ段　＊　にくい)
; (接尾辞　形容詞性述語接尾辞　イ形容詞アウオ段　＊　づらい)))

例えば (2) は、動詞基本連用形に形容詞性述語接尾辞が連接する例である。この規則に上記のような制限を設けることによって、SFJの7課で学ぶ「書きたい」のように動詞基本連用形「書き」に「たい」が続く文字列は解析するが、「かかせたい・かきやすい・かきにくい・かきづらい」のようにSFJでは出てこない表現は、他の可能性がない場合[5]、未定義語として出力される。

5．評価方法

5．1　評価の観点

本研究では、システム評価のために、情報検索の分野で一般的に用いられている再現率 (recall) と適合率 (precision) という観点を使用する。再現率とは、全ての誤りのうち、システムが未定義語として指摘したものの割合であり、正しく誤りを指摘できた割合を指すこととな

る。適合率とは、指摘された未定義語のうち、実際に誤りであった割合を表し、適合率が高ければ、ノイズ（正しい部分なのに誤って未定義語として指摘されたもの）が少なく、逆に適合率が低ければ、ノイズが多いこととなる。

$$(3)\ \text{再現率} = \frac{\text{指摘した誤り}}{\text{全ての誤り}} \qquad \text{適合率} = \frac{\text{指摘した誤り}}{\text{指摘された未定義語}}$$

再現率と適合率のどちらを重視するかに関してであるが、誤りを見逃すよりは、誤りを含む可能性を多めに表示させることの方が望ましいと考え、本研究では再現率をより重視する[(6)]。

5．2　誤りの定義

実験に使用したデータに対し、まず手作業により誤りとする部分の特定を実施した。ここで、誤りの定義について言及する。本研究では、表記・連接のチェックを行うシステムの検証を行うため、表記および語と語の連接がおかしいと思われるものを誤りとする。先行研究で指摘されているように、日本語学習者が犯す誤りにはいろいろなものがあるが[(7)]、ここで特定する誤りは、本稿で提案する日本語チェッカーが指摘すべき誤りである。（以後、下線部を誤りとする）

5．2．1　誤りとするもの

ここで正しく表記されていないと考えるものには、語の一部が抜けていたり、濁点が抜けているもの、逆に余分につけてしまったものなどがある。例えば、「留学がいた」「かそく（家族の誤り）」などで、また「少少」やカタカナ語の誤表記も含む。また中国語からの影響と思われるもので、「郵電局に」「生活の楽趣」などのように、日本語にはない語彙が使われている場合も誤りと考える。また、後述するように一部例外もあるが、活用に関する誤りもここに含める。例えば、「と思う」の前は用言の基本形（plain form）が出現するので、それ以外の「静かと思います」のような表現は誤りとする。

また品詞を間違えて使用している場合も含む。これは「じんせい」を形容動詞として使用している「じんせいなとき」のような場合である。

5．2．2　誤りとしないもの

次に、誤りとしないものについて説明する。まず基本的な方針として、意味的処理が必要なものは、本システムが指摘すべき誤りからは除外する。本来の作文システムでは必須と考えられるが、本システムの目的は3．2で示した通り、意味処理のための前処理であるので、この部分の誤りは除外した。例えば「ときの日が」は意味的におかしいが、＜名詞＞の＜名詞＞という形でつながっており連接としておかしくないものである。このように連接規則で排除できないものは誤りとしない。他に誤りとしないものには、ハとガの選択の誤り、動詞選択の誤り、テンスの誤りがある。動詞選択の誤りには、例えば「あげました」と「くれました」や、「泊まりました」と「泊めました」等が含まれる。

また、名詞選択の誤りに関しても、意味的な処理が必要となるものは除外した。例えば、「はじめ」「はじめは」「はじめて」のように意味的要素が関わっているもので、使い分けが間違っている場合は誤りとしないが、「はじめで」は12課までに含まれる用法ではないので、誤りとした[8]。このように誤りの認定においては学習者がどの課まで進んでいるかに大きく依存することになる。

さらに、助詞の使用における誤りに関しては、本システムが指摘すべき誤りから除外する。助詞の使用が正しいかどうかは、格文法などを用いて名詞と動詞の関係から判定しないと明らかにならないためである。なお、3．2で示したように、連接規則から導かれるある形式に連接する形式がない場合、未定義語となるため、「おじいさんプエルトリコかえります」のように助詞が抜けている部分や、「上海でより」「日本ののたべもの」のように不要な部分に助詞が入っている場合は、本システムが指摘すべき誤りとする。

また文末に「だ・です」が混在して使われている丁寧さの不統一や、人称とモダリティが一致していない表現も誤りとしない。これは、係り受けなどと同様に、隣り合っていないものが関係する場合、何らかの情報を保持しておくスタック処理が要求されるもので、この種の処理は本システムでは行っていないためである。

また、意味処理が加わらないため、活用の誤りと思われるものでも、誤りを含む形が別の動詞として解釈できる場合には、誤りとしない。例えば「ひこうきにのんで」のような文で、「のんで」は「のって」の誤りと考えられるが、この場合「9時にのんで」が適格とされるために、本システムでは誤りとはしない。

6．システム評価

6．1　比較対象

本論文で提案する手法の有効性を検証するために、以下の4つのシステムに作文を入力した際のチェック結果に対して、再現率と適合率を割り出し、比較・評価を行う。

a．茶筌をそのまま使用
b．辞書を12課までの語彙に制限し、連接規則はそのまま使用
c．辞書は茶筌をそのまま使用し、連接規則を12課までの文法に制限
d．辞書、連接規則を12課までの語彙・文法に制限

6．2　カウント方法

まず、どのような場合、誤りが指摘できたと考えるかについて、方針と具体例を示す。本システムの目的は、入力した文に対して、下線の付いている部分（またはその前後）に学習者の注意を向けさせることにあるので、注意を促すべき対象、またその前後にシステムの指摘が現れるかどうかがカウントの基準となる。これから、具体例を3つに分けて説明する。

6．2．1　完全一致

まず筆者の指摘する誤りと、システムが指摘するものが、完全に一致する場合がある。以後「筆」を筆者が指摘する誤りの箇所、「シ」をシステムが指摘する誤りの箇所とする。例えば、「レストーラン（筆）」と「レストーラン（シ）」や「言ました（筆）」と「言ました（シ）」、「いっしょう（筆：いっしょ）」と「いっしょう（シ）」のような場合である。

6．2．2　部分一致

筆者が指摘する部分と、システムが指摘する部分にずれが生じているが、誤りが指摘できているとしてカウントするものには、以下のような場合がある。同じ「にゅう学日」という文字列に対して、筆者は、「にゅう学日」という語はないので「にゅう学日」を誤りと指摘したが、システムは「にゅう学日」を誤りと指摘する。なぜなら「に」は数字の２として解析するが、これに続く語はないので、これ以降の部分を未定義語として指摘している。このような場合も、指摘されたとしてカウントする。

また、一部が重なっていればカウントする場合もある。例えば、筆者は「いえがをみて」を「えいが」の誤りと判断して「いえが」を誤りとした。システムはこの文字列に対して、「いえ」を「家」と解析するが、それに後接するものでさらに「を」に続くものがないため、「いえがをみて」を誤りとして指摘している。このような部分一致には、他に「帰えりました（筆）」と「帰えりました（シ）」や、「がんがえりました（筆）」と「がんがえりました（シ）」などのような例もある。この場合システムが指摘した部分は共に誤りの一部なので、誤りが指摘されたとして適合率の集計の際、２つとしてカウントする[9]。

6．2．3　不一致

筆者の指摘する誤りと、システムが指摘する部分が一致しない場合がある。これは誤りが含まれる位置が文頭の時に起こると考えられるが、このような場合もカウントするケースがある。例えば「たくさん20分ぐらい（筆）」と「たくさん20分ぐらい（シ）」の場合、「たくさん」が間違っていると考えられるが、システムは連接規則から文頭に位置する「たくさん」に「20」という数詞が連接することはないと考え、20の部分を未定義語と検出する。この場合も、学習者に誤りとして指摘した部分の前後に注意を促すことができたとしてカウントする。以上の方針で誤りをカウントする。

6．3　結果の検討

作文に対してそれぞれのシステムが未定義語と判定した部分に対して、誤りの集計を行った結果が表１である。

表１．再現率・適合率

辞書・連接規則	再現率	指摘した誤り／全ての誤り	適合率	指摘した誤り／指摘された未定義語
a. 茶筌・茶筌	20.6%	65/316	66.7%	68/102
b. 12茶・茶筌	63.6%	201/316	38.8%	213/549
c. 茶筌・12茶	25.0%	79/316	61.9%	83/134
d. 12茶・12茶	75.9%	240/316	40.3%	252/625

この結果を見ると、再現率は辞書・連接規則共に12茶である場合（d）が最も高く、適合率は辞書・連接規則共に茶筌のもの（a）が最も高いことがわかる。誤りを探すという本来の趣旨に照らすと、指摘した誤りの数が最も多かったのはdであり、本研究で提案した、語彙・連接規則を制限するという考え方に基づく表記チェッカーの有効性が示されたといえよう。以後、結果について詳細に考察を加える。

６．３．１　全体の考察

まず、辞書に茶筌を使用した場合（a・c）、辞書に12茶を使用したb・dに比べて、共に再現率が低い。これは、豊富な辞書の中から古語や固有名詞・地名など様々な可能性を探し出すため、未定義語として出力されるものが少ないためである。例えば、「みがいて」の誤った形と思われる「みかって」は「みか（固有名詞）」「って（副助詞）」と解析される。また、a・cで指摘した誤りの数は、b・dで指摘した誤りの数の３分の１ほどでしかない。このように辞書に茶筌を使用した場合、検出数自体が少なく、再現率もきわめて低いといえる。逆に辞書に12茶を使用した場合（b・d）、a・cに比べると、指摘した未定義語はa・bで５倍強（a＝102、b＝549）、c・dでも５倍近い（c＝134、d＝625）数が指摘された。つまりノイズをより多く拾っていることになるが、再現率は高い。再現率を重視する立場なら、辞書に12茶を利用した方が良い結果をもたらすといえる。

次に、連接規則を中心に見てみる。連接規則に茶筌を用いたa・bの場合と、12茶を用いたc・dの場合を比べてみる。前述したとおり、辞書に何を用いるかで、再現率が大きく変わっているが、それを考慮すれば、連接規則に茶筌を用いたa・bよりも、12茶を用いたc・dの方が、いずれも再現率は良い。ただ、その違いは辞書の違いによってもたらされる影響ほど強くはないようである[(10)]。

以上から、辞書の影響が大きく、辞書にどれを用いるかが結果に大きく関係しているといえる。そして連接規則は辞書ほどではないが、結果に少なからず影響していることが指摘できる。

６．３．２　dのシステムに関する考察

ここで、本研究で提案した考えを具体化したd（辞書・12茶、連接規則・12茶）について詳

細に検討を加えていく。この場合、再現率が最も高く、誤りを多く収集していることがわかるが、反面、未定義語の収集が多くなり適合率を下げる結果となった。そこで収集した未定義語に対して詳細な検討を加え、課題を明らかにしていく。

誤って未定義語と指摘される例を調査したところ、次のような問題があることが分かった。一番多い未定義語は、「幸せな」と「生」であり、「生」は「人生」の一部として現れている。これは、本研究で対象とした作文のうち、1学期分の作文のタイトルが「人生で一番幸せな日」で、「人生」「幸せ」共に12課までの語彙ではなかったことによる(11)。他に正しく書けているのに未定義語とされてしまう例として、「作られる所」「だんだん慣れました」など12課以降で学習する表現がある。作文を収集した日本語150には、日本語が未習の学生や既習の学生が含まれ、そのため作文には、このようにかなりバラエティに富んだ語彙や文型が使われており、その表現の多さが適合率を下げる結果になっている。このクラスで日本語学習を全くの初級から開始した学生の作文でないことは、システムの厳密な検証、すなわち評価テストに使用するデータの質という意味で課題を残すものである。

また、作文に出現する生活語彙の問題が見られる。作文中には、「マクドナルド」、「バー」、また個人名など、教科書に未出の語彙が見られる。例えば、教科書に出てくる「木村先生」は辞書に登録されているが、それ以外の実在する教官の名前が作文が出た場合、誤り候補として検出される可能性が高い。語彙の豊富な学生の作文に多く見られることであるが、生活語彙が多く使われた作文は、全てが未定義語とされるため、適合率が低くなってしまうので、どういった対策をとるべきかが次の課題となる。また、生活語彙を登録するにあたり、どのようなデータを参考にするのがいいかについては、アドホックにではなく、規則的にモデルを選択するということを考えていく必要があろう。

その他に、パラメーター設定の問題がある。茶筌では連接規則のコスト（重み付け）も使用者側で設定することが可能である。しかし本研究では、コストに関しては変更を加えていない。そのため、例えば句読点の使い方に関して、作文のチェックに厳密な連接規則が適用されるため、(4) のように引用文の中についた句点は未定義語となる。

(4) 授業休んでもよろしでしょうか。」と言ました。

この問題については、コストを変更することで、このような部分も適格と判断できるようになる。ただこれには、日本語学習者の書く作文にはどういったものを求め、期待するかという、一つの方針が要求されるだろう。同時に作文のチェックとして正誤の判断基準をどう定めていくか、さらに厳密な検討が必要となる。また本研究では、誤りである可能性をより多く持つと考えられる未定義語の検出を重視するという観点から、入力された作文が正しく形態素解析されているかという点については、現段階では十分に考慮されていない。さらに、本研究で使用

したチェッカーの改良にあたっては、日本語学習者の作文のチェックに適した連接規則のコストをどのように設定すべきかという問題も考慮する必要がある。茶筌では、品詞のコストづけも設定可能であるので、日本語チェッカーとして最適な連接規則・品詞のコストを探ることも必要であろう。

本研究をまとめると、結果として、辞書に12茶を使い、連接規則に12茶を使うdが最良であること、すなわち語彙・文法規則を制限したシステムを用いた場合が最も誤りの可能性を多く指摘しうることが明らかになり、3．1で述べたシステム作成の前提は正しいことが検証された。このことは自由作文をも対象とした十分実用的なシステムが構築可能であることを意味している。しかし、適合率の上昇を目指した場合、学習者がふだん接している生活語彙の適切な登録が必要となり、ベストと思われる品詞と連接規則のコストの設定など課題も残されている。

7．今後の予定

本稿のシステムは、3．1で述べたように文法診断システムの前処理的位置づけとして想定している。本システムが対応しない点として挙げた意味情報が必要な表現、つまり文法上の間違い、ガとハの選択に関する問題、格構造のチェックなどは、表記チェッカー以降のシステムで対応することになる。今後、文法診断システムを開発するにあたっては、システムの構築とともに、表記チェッカー自体の精度を一層高めることが不可欠となる。

また、チェッカー自体の問題として、厳密に学習した課までの語彙と文法でいいのか、検証する必要がある。12課までの学習者に対して、単純に12課までの茶筌（12茶）を用意したが、望ましい結果を得るためには、範囲をどうするか考慮する必要がある。例えば「絵」という単語は、13課で出てくるため、12茶では未定義語とされてしまう。チェッカーで使用する辞書の設定として、多少幅を持たせることも可能であるし、また、SFJのテキストは3巻構成なので各巻ごとに分けるという分類も考えられ、どういった分け方がチェッカーとして望ましいのか検討する余地がある。

最後に、本稿で提案したシステムは、視点を変えれば、作文中にどの課の語彙がどのくらい使われているかという方向での出力も可能になる。これは作文の評価などに関連すると思われ、今後、作文評価システムの構築を視野に入れて、データ分析・検討も行なう予定である。

注

(1) URL：http：//chasen.aist-nara.ac.jp. 本研究ではver1.0を使用した。執筆現在は2.2.8にバージョンアップされているが、変更点は辞書・連接規則の改良であり、本研究では、独自に辞書・連接情報を用意するため、特に新しいバージョンである必要がないと考えられるので、本研究では1.0のまま使用した。

(2) 筑波ランゲージグループ（1991）『Situational Functional Japanese』凡人社

(3) 具体的にはSFJ DRILLS内のNew Words in DrillsとAdditional New Words in Drills、NOTES内のWords in the conversationとWords in the reportから抜き出した。

(4) この用語は、松本他（1997）の使用説明書によった。

(5) 例えば、「かきやすい」には「かきやすいかをたべた」（柿やスイカ）の一部となるような場合が考えられるが、「柿」も「スイカ」も12茶では未定義語である。

(6) 表示されるものに、ノイズが多く含まれれば、それだけ学習者に混乱を与えかねないという指摘も考えられる。しかし、ノイズをゼロにすることは現段階では不可能であるし、学習者により多くの考えるきっかけを与えることが重要であると考える。また表記に関しては、この段階でなるべく多くの誤りを訂正する機会を与えることで、次に想定する文法診断システムへの負担を軽減する目的もある。

(7) 森田（1985）、大曽（1986、87）、田窪（1987）など。

(8) 例えば、「はじめて、私は日本語がぜんぜんできなかったから、じゅぎょうがこわかった」「はじめ日本に行きました」のような場合、誤りとしないが、「はじめで私のだけかのじょをすきですが……」は誤りとする。

(9) これは、例えば「がんがえりました（シ）」の場合、システムは未定義語を2つとカウントしているので、指摘された誤りは2つとする。

(10) ただし、適合率から見れば、aよりcで数値が上がるはずのものが、逆転している。これについては、辞書が茶筌で豊富なために、12茶で規則を制限してしまうと、それに適応した規則がないため、未定義語が多くなると考えられるし、12茶では未定義語とした固有名詞を茶筌では正しく解析したためと考えられる。

(11) 全体の母集合に対する比率にもよると思われるが、「生」と「幸」が一番多かったことについては、本論で使用した作文が、システムの評価テストの対象として適切性の面で問題がないとはいえない。

本研究は、日本学術振興会科学研究費補助金基盤研究（C）（2）課題番号12680297（研究代表者　西村よしみ）の助成を受けている。

参考文献

大曽美恵子（1986、87）「誤用分析（1）～（6）」『日本語学』5-10～6-2

掛川淳一・神田久幸・藤岡英太郎・伊丹誠・伊藤紘二（2000）『日本語学習支援システムにおける作文診断処理系の提案と試作』、電子情報通信学会論文誌D-I Vol. J83-D-I、No.6、pp.693-701

田窪行則（1987）「誤用分析（1）～（6）」『日本語学』6-4～6-10

松本裕治・北内啓・山下達雄・平野善隆・今一修・今村友明（1997）日本語形態素解析システム『茶筌』version1.0使用説明書 NAIST Technical Report NAIST-IS-TR97007

森田良行（1985）『誤用文の分析と研究—日本語学への提言—』明治書院
山本幹雄・森輝彦（2000）『日本語作文学習システムのための統計的言語モデルを用いた日本語チェッカー』科学研究費特定領域研究（A）「メディア教育利用」A02「外国語教育の高度化の研究」領域全体会議資料